Probability in the Engineering and Informational Sciences, **19**, 2005, 83–97. Printed in the U.S.A.

SIZE-BIASED PERMUTATION OF DIRICHLET PARTITIONS AND SEARCH-COST DISTRIBUTION

Javiera Barrera

Departmento Ingeniería Matemática and Centro Modelamiento Matemático
UMR 2071 UCHILE-CNRS
Santiago, Chile
E-mail: barrera@dim.uchile.cl

Thierry Huillet

Laboratoire de Physique Théorique et Modélisation
CNRS-UMR 8089 et Université de Cergy-Pontoise
Neuville sur Oise, France
E-mail: huillet@ptm.u-cergy.fr

Christian Paroissin

MODAL'X
Université de Paris 10 Nanterre
92001 Nanterre Cédex, France
E-mail: cparoiss@u-paris10.fr

Consider the random Dirichlet partition of the interval into n fragments at temperature $\theta > 0$. Explicit results on the law of its size-biased permutation are first supplied. Using these, new results on the comparative search cost distributions from Dirichlet partition and from its size-biased permutation are obtained.

1. INTRODUCTION AND DESCRIPTION OF MAIN RESULTS

Basic facts on the random Dirichlet partition of the interval into n fragments at temperature $\theta > 0$ are first recalled in Section 2. In Section 3, explicit results on the law of its size-biased permutation are supplied. A size-biased permutation of the fragments sizes is the one obtained in a size-biased sampling process without

https://doi.org/10.1017/S0269964805050060 Published online by Cambridge University Press

replacement from a Dirichlet partition. The main points that we develop are the following: In Proposition 1, it is recalled that the length of an interval containing a random sample is stochastically larger than the typical fragment size from a Dirichlet distribution. Its law is computed and the stochastic domination result is made more explicit in Corollary 2. In Theorem 3, the law of the length of the kth fragment in the size-biased permutation is supplied. It is also shown there that the consecutive fragments in the size-biased permutation are arranged in stochastic descending order. In Corollary 4, the expected length of the kth fragment in the size-biased permutation is supplied. In Theorem 5, we give the joint law of the size-biased permutation fragments sizes explicitly (or, rather, its joint moment function).

Size-biased permutations of random discrete distributions are known to be the random equilibrium distributions of the heaps process consisting in moving sequentially sampled fragments to the front, starting from the original partition. Using the computations from Section 3, new results on the comparative search-cost distributions from Dirichlet partition and from its size-biased permutation are obtained in Section 4. The search cost of an item in a library is the number of items above it in the heap; averaging over the items gives the search cost of a typical item. The search cost when the library has reached the equilibrium state is expected to be smaller than the search cost in the original Dirichlet partition itself. The results that we describe confirm this intuition. In Proposition 6, the limiting search cost per item in a Dirichlet partition is first shown to be uniformly distributed. In Lemma 7, we compute explicitly the law of the search cost in the size-biased permutation of a Dirichlet partition, using Corollary 4. First and second moments are also obtained differently from the techniques usually employed to do so. In Theorem 8, the limiting search cost per item in a size-biased permutation of the Dirichlet partition is shown to be beta$(1,1 + 1/\theta)$ distributed. Finally, in Proposition 9, considering the asymptotic introduced by Kingman, $n \uparrow \infty$, $\theta \downarrow 0$, $n\theta = \gamma > 0$, we find the limiting size-biased permutation search cost to be geometrically distributed.

2. PRELIMINARIES: THE DIRICHLET DISTRIBUTION $D_n(\theta)$

We will consider the following random partition into n fragments of the unit interval: Let $\theta > 0$ be some parameter that we will interpret as temperature or disorder of the partition. Assume that the random fragments' sizes $\mathbf{S}_n := (S_1, \ldots, S_n)$ (with $\sum_{m=1}^{n} S_m = 1$) is distributed according to the (exchangeable) Dirichlet $D_n(\theta)$ density function on the simplex; that is to say,

$$f_{S_1,\ldots,S_n}(s_1,\ldots,s_n) = \frac{\Gamma(n\theta)}{\Gamma(\theta)^n} \prod_{m=1}^{n} s_m^{\theta-1} \cdot \delta_{\left(\sum_{m=1}^{n} s_m - 1\right)}. \tag{1}$$

Alternatively, the law of $\mathbf{S}_n := (S_1, \ldots, S_n)$ is characterized by its joint moment function

$$\mathbf{E}\left[\prod_{m=1}^{n} S_m^{q_m}\right] = \frac{\Gamma(n\theta)}{\Gamma\left(n\theta + \sum_{m=1}^{n} q_m\right)} \prod_{m=1}^{n} \frac{\Gamma(\theta + q_m)}{\Gamma(\theta)}. \tag{2}$$

https://doi.org/10.1017/S0269964805050060 Published online by Cambridge University Press

In this case, $S_m \overset{d}{=} S_n$, $m = 1, \ldots, n$, independently of m and the individual fragment sizes are all identically distributed (i.d.). Their common density on the interval (0,1) is given by

$$f_{S_n}(s) = \frac{\Gamma(n\theta)}{\Gamma(\theta)\Gamma((n-1)\theta)}\, s^{\theta-1}(1-s)^{(n-1)\theta-1}, \qquad \textbf{(3)}$$

which is a beta$(\theta, (n-1)\theta)$ density, with mean value $\mathbf{E}(S_n) = 1/n$ and variance $\boldsymbol{\sigma}^2(S_n) = (n-1)/[n^2(n\theta+1)]$.

We recall that a random variable, say $B_{a,b}$, with $B_{a,b} \overset{d}{\sim}$ beta(a,b), has density function $f_{B_{a,b}}(x) := [\Gamma(a+b)/\Gamma(a)\Gamma(b)]x^{a-1}(1-x)^{b-1}$, $a, b > 0$, $x \in [0,1]$ and moment function $\mathbf{E}[B_{a,b}^q] = [\Gamma(a+q)\Gamma(a+b)]/[\Gamma(a)\Gamma(a+b+q)]$, with $\Gamma(a)$ the Euler's Gamma function.

We also recall that when $\theta = 1$, the partition model [Eqs. (1) and (2)] corresponds to the standard uniform partition model of the interval.

From Eq. (3), as $n \uparrow \infty$, we next have

$$nS_n \xrightarrow{d} \Gamma_{\theta,\theta}, \quad \text{with density } f_{\Gamma_{\theta,\theta}}(t) = \frac{\theta^\theta}{\Gamma(\theta)}\, t^{\theta-1}e^{-\theta t}, \qquad t > 0, \qquad \textbf{(4)}$$

showing that the sizes of fragments are asymptotically all of order $1/n$.

Consider next the sequence $\mathbf{S}_{(n)} := (S_{(m)}; m = 1, \ldots, n)$ obtained while ranking the fragment sizes $\mathbf{S}_n$ according to descending sizes, hence with $S_{(1)} > \cdots > S_{(m)} > \cdots > S_{(n)}$. The $S_{(m)}$ distribution can hardly be derived in closed form. However, one could prove that as $n \uparrow \infty$,

$$n^{(1+\theta)/\theta} S_{(n)} \xrightarrow{d} W_\theta \quad \text{and} \quad n\theta\left(S_{(1)} - \frac{1}{n\theta}\log\left(n(\log n)^{\theta-1}\right)\right) \xrightarrow{d} G_\theta, \qquad \textbf{(5)}$$

where W_θ is a Weibull random variable and G_θ is a Gumbel random variable such that $\mathbf{P}(W_\theta > t) = \exp[-t^\theta/s_\theta]$, $t > 0$, and $\mathbf{P}(G_\theta \leq t) = \exp[-s_\theta^{-1}\exp(-t)]$, $t \in \mathbb{R}$, $s_\theta := \Gamma(1+\theta)\theta^{-\theta} > 0$ is a scale parameter.

In the random division of the interval as in Eq. (1) at disorder θ, although all fragments are identically distributed with sizes of order n^{-1}, the smallest fragment's size grows like $n^{-(\theta+1)/\theta}$ and the largest is of order $(1/n\theta)\log(n(\log n)^{\theta-1})$. The smaller θ is, the larger (smaller) the largest (smallest) fragment's size is; hence, the smaller disorder θ is, the more the values of the S_m are, with high probability, disparate. At low disorder, the size of the largest fragment $S_{(1)}$ tends to dominate the other ones and the range $S_{(1)} - S_{(n)}$ increases when θ decreases.

To the contrary, large values of θ correspond to situations in which the range of fragments' sizes is lower: the fragments' sizes look more homogeneous and distribution equation (1) concentrates on its center. At high disorder, the diversity of the partition is large.

3. SAMPLING WITHOUT REPLACEMENT AND SIZE-BIASED PERMUTATION OF THE FRAGMENTS

Assume some observer is sampling such an interval as follows: Drop at random points onto this randomly broken interval and record the corresponding numbers of visited fragments. Consider the problem of determining the order in which the various fragments will be discovered in such a sampling process. To avoid revisiting many times the same fragment once it has been discovered, we need to remove it from the population as soon as it has been met in the sampling process. However, to do that, an estimation of its size is needed. We first do that for the first visited fragment. Once this is done, after renormalizing the remaining fragments' sizes, we are left with a population of $n-1$ fragments, the sampling of which will necessarily supply a so far undiscovered fragment. Its size can be estimated and so forth, renormalizing again, until the whole available fragments' population has been visited. In this way, not only can the visiting order of the different fragments be understood but also their sizes. The purpose of this section is to describe the statistical structure of the size-biased permutation of the fragments' sizes as those obtained while avoiding the ones previously encountered in a sampling process.

Let $\mathbf{S}_n := (S_1,\ldots,S_n)$ be the random partition of the interval $[0,1]$ considered here, with $S_m \overset{d}{=} S_n \overset{d}{\sim} \mathrm{beta}(\theta,(n-1)\theta)$, $m = 1,\ldots,n$, $\sum_m S_m = 1$.

Let U be a uniformly distributed random throw on $[0,1]$ and let $\mathfrak{L}_n := \mathfrak{L}_n(U)$ be the length of the interval of the random partition containing U. The distribution of $\mathfrak{L}_n$ is characterized by the conditional probability

$$\mathbf{P}(\mathfrak{L}_n = S_m | \mathbf{S}_n) = S_m. \tag{6}$$

In this size-biased picking procedure, long intervals are favored and one expects that $\mathfrak{L}_n \succcurlyeq S_n$ in the usual stochastic sense that $\bar{F}_{\mathfrak{L}_n}(s) \geq \bar{F}_{S_n}(s)$, $\forall s \in [0,1]$.

Let us first check that the size of the interval containing U is stochastically larger than the typical fragment's length of the original partition.

3.1. Length of the First Size-Biased Sample

From the size-biased picking construction, it follows (see, e.g., [6]) that for all nonnegative measurable functions φ on $[0,1]$,

$$\begin{aligned} \mathbf{E}[\varphi(\mathfrak{L}_n)/\mathfrak{L}_n] &= \mathbf{E}[\mathbf{E}[\varphi(\mathfrak{L}_n)/\mathfrak{L}_n | \mathbf{S}_n]] \\ &= \mathbf{E}\left[\sum_{m=1}^{n} \varphi(S_m)/S_m \mathbf{P}(\mathfrak{L}_n = S_m | \mathbf{S}_n)\right] \\ &= \mathbf{E}\left[\sum_{m=1}^{n} \varphi(S_m)\right]. \end{aligned} \tag{7}$$

Taking in particular $\varphi(x) = x\mathbf{I}(x > s)$ in Eq. (7), we get

$$\bar{F}_{\mathfrak{L}_n}(s) = \mathbf{E}\left[\sum_{m=1}^{n} S_m \mathbf{I}(S_m > s)\right],$$

which, since $S_m \overset{d}{=} S_n$, $m = 1, \ldots, n$, is

$$\bar{F}_{\mathfrak{L}_n}(s) = \sum_{m=1}^{n} \int_s^1 t\, dF_{S_m}(t) = n \int_s^1 t\, dF_{S_n}(t). \tag{8}$$

Proposition 1: $\mathfrak{L}_n \overset{d}{\sim}$ *beta*$(1+\theta,(n-1)\theta)$ *and it holds that*

$$\mathfrak{L}_n \succcurlyeq S_n. \tag{9}$$

Proof: The condition $\bar{F}_{\mathfrak{L}_n}(s) \geq \bar{F}_{S_n}(s)$ holds for all s in $[0,1]$ because this is equivalent to $\int_s^1 t\, dF_{S_n}(t)/\bar{F}_{S_n}(s) \geq \mathbf{E}(S_n)$, which is always true because the left-hand side is the conditional expectation of S_n given $S_n > s$, certainly larger than $\mathbf{E}(S_n)$ itself. Because $S_n \overset{d}{\sim}$ beta$(\theta,(n-1)\theta)$, one can check directly that $\mathfrak{L}_n \overset{d}{\sim}$ beta$(1+\theta, (n-1)\theta)$, with $\mathbf{E}(\mathfrak{L}_n) = (1+\theta)/(n\theta+1)$. ■

This apparent paradox (discussed when $\theta = 1$ in Feller [8, pp. 22–23] and subsequently worked out in Hawkes [12, pp. 294–295]) may be understood by observing that in the size-biased picking procedure, long intervals are favored. It constitutes the version on the interval of the standard waiting-time paradox on the half-line. As a corollary, the following decomposition holds.

Corollary 2: *Let B_n be a Bernoulli random variable with parameter $1/n$ and $B_{\theta,1} \overset{d}{\sim}$ beta$(\theta,1)$ on $[0,1]$, independent of B_n. Define a $[0,1]$-valued random variable R_n with distribution*

$$R_n \overset{d}{=} B_n + (1-B_n) \cdot B_{\theta,1}. \tag{10}$$

Then, the following decomposition holds:

$$R_n \cdot \mathfrak{L}_n \overset{d}{=} S_n, \tag{11}$$

where R_n and $\mathfrak{L}_n$ are independent.

Proof: Since $\mathbf{P}(B_n = 1) = 1/n$, we have

$$\mathbf{E}[R_n^q] = \frac{1}{n} + \left(1 - \frac{1}{n}\right) \frac{\theta}{\theta + q}.$$

Taking $\varphi(x) = x^{q+1}$ in Eq. (7), the moment function of $\mathfrak{L}_n$ reads $(q > -(1+\theta))$

$$\mathbf{E}[\mathfrak{L}_n^q] = \mathbf{E}\left[\sum_{m=1}^{n} S_m^{q+1}\right] = n\mathbf{E}[S_n^{q+1}] = \frac{n\Gamma(n\theta)\Gamma(\theta+q+1)}{\Gamma(\theta)\Gamma(n\theta+q+1)},$$

recalling that $\mathbf{E}[S_n^q] = [\Gamma(n\theta)\Gamma(\theta+q)]/[\Gamma(\theta)\Gamma(n\theta+q)]$ is the common moment function of S_m, $m = 1, \ldots, n$, with $\mathbf{E}(S_n) = 1/n$. So,

$$\mathbf{E}[S_n^q] = \frac{n\theta + q}{n(\theta + q)} \mathbf{E}[\mathfrak{L}_n^q] = \mathbf{E}[R_n^q]\mathbf{E}[\mathfrak{L}_n^q]. \qquad ■$$

https://doi.org/10.1017/S0269964805050060 Published online by Cambridge University Press

3.2. Size-Biased Permutation of the Fragments

Consider the random partition $\mathbf{S}_n$. Let $L_1 := \mathfrak{L}_n$ be the length of the first randomly chosen fragment $M_1 := M$, so with $L_1 := S_{M_1}$ and $\mathbf{P}(M_1 = m_1 | \mathbf{S}_n) = S_{m_1}$. A standard problem is to iterate the size-biased picking procedure, by avoiding the fragments already encountered: By doing so, a size-biased permutation (SBP) of the fragments is obtained. We study here this process in some detail.

In the first step of this size-biased picking procedure,

$$\mathbf{S}_n := \mathbf{S}_n^{(0)} \to (L_1, S_1, \ldots, S_{M_1-1}, S_{M_1+1}, \ldots, S_n),$$

which can be written as $\mathbf{S}_n \to (L_1, (1 - L_1)\mathbf{S}_{n-1}^{(1)})$, with

$$\mathbf{S}_{n-1}^{(1)} := (S_1^{(1)}, \ldots, S_{M_1-1}^{(1)}, S_{M_1+1}^{(1)}, \ldots, S_n^{(1)}),$$

a new random partition of the unit interval into $n-1$ random fragments.

Given $L_1 \overset{d}{\sim} \text{beta}(1+\theta, (n-1)\theta)$, the conditional joint distribution of the remaining components of $\mathbf{S}_n$ is the same as that of $(1 - L_1)\mathbf{S}_{n-1}^{(1)}$, where the $(n-1)$-vector $\mathbf{S}_{n-1}^{(1)} \overset{d}{\sim} D_{n-1}(\theta)$ has the distribution of a Dirichlet random partition into $n-1$ fragments. Next, pick at random an interval in $\mathbf{S}_{n-1}^{(1)}$ and call V_2 its length, now with distribution $\text{beta}(1+\theta, (n-2)\theta)$, and iterate until all fragments have been exhausted.

With $V_1 := L_1$, the length of the second fragment by avoiding the first reads $L_2 = (1 - V_1)V_2$. Iterating, the final size-biased permutation (SBP) of $\mathbf{S}_n$ is $\mathbf{L}_n := (L_1, \ldots, L_n)$. We will set $\mathbf{L}_n = \text{SBP}(\mathbf{S}_n)$.

From this construction, if $(V_1, \ldots, V_{n-1})$ is an independent sample with distribution $V_k \overset{d}{\sim} \text{beta}(1+\theta, (n-k)\theta)$, $k = 1, \ldots, n-1$, then,

$$L_k = \prod_{i=1}^{k-1} (1 - V_i)V_k, \qquad k = 1, \ldots, n-1, \tag{12}$$

$$L_n = 1 - \sum_{k=1}^{n-1} L_k = \prod_{k=1}^{n-1} (1 - V_i) \tag{13}$$

is the stick-breaking scheme construction of the size-biased permutation of $\mathbf{S}_n$. Note that $\overline{V}_i := 1 - V_i \overset{d}{\sim} \text{beta}((n-i)\theta, 1+\theta)$ and that V_n should be set to one. From this well-known construction and properties (see Kingman [16, Chap. 9, 9.6], Patil and Taillie [17], and Donnelly [4]) we obtain that the L_k's, $k = 1, \ldots, n$, are arranged in stochastically decreasing order. More precisely, we have the following:

THEOREM 3:

(i) *The law of L_k, for $k = 1, \ldots, n$, is characterized by*

$$\begin{aligned}
\mathbf{E}[L_k^q] &= \prod_{i=1}^{k-1} \mathbf{E}[\overline{V}_i^q]\mathbf{E}[V_k^q] \\
&= \prod_{i=1}^{k-1} \frac{\Gamma((n-i)\theta + q)\Gamma((n-i+1)\theta + 1)}{\Gamma((n-i)\theta)\Gamma((n-i+1)\theta + 1 + q)} \\
&\quad \times \frac{\Gamma(1+\theta+q)\Gamma(1+(n-k+1)\theta)}{\Gamma(1+\theta)\Gamma(1+(n-k+1)\theta+q)}.
\end{aligned} \tag{14}$$

(ii) *Let* $B_{(n-k+1)\theta,1} \overset{d}{\sim}$ *beta*$((n-k+1)\theta,1)$. *Then,*

$$L_k \overset{d}{=} B_{(n-k+1)\theta,1} \cdot L_{k-1}, \qquad k = 2,\ldots,n, \tag{15}$$

where pairs $B_{(n-k+1)\theta,1}$ *and* L_{k-1} *are mutually independent for* $k = 2,\ldots,n$.

(iii) $L_1 \geqslant \cdots \geqslant L_k \geqslant \cdots \geqslant L_n$.

PROOF:

Part (i) is a direct consequence of the construction, since $\overline{V}_i := 1 - V_i \overset{d}{\sim}$ beta$((n-i)\theta,1+\theta)$, $i = 1,\ldots,k-1$, and $V_k \overset{d}{\sim}$ beta$(1+\theta,(n-k)\theta)$ are mutually independent. Recalling the expression of the moment function for beta distributions, the corresponding expression of $\mathbf{E}[L_k^q]$ follows.

Part (iii) being clearly a consequence of (ii), it remains to prove (ii).

Regrouping terms directly from Eq. (14), we have $\mathbf{E}[L_k^q] = \mathbf{E}[L_{k-1}^q]\mathbf{E}[B_k^q]$, with

$$\mathbf{E}[B_k^q] = \frac{\Gamma((n-k+1)\theta+q)}{\Gamma((n-k+1)\theta)} \frac{\Gamma(1+(n-k+1)\theta)}{\Gamma(1+(n-k+1)\theta+q)}.$$

This is the moment function of a beta$((n-k+1)\theta,1)$-distributed random variable. ■

Result (ii) is also in Collet, Huillet, and Martinez [3], with a slightly different proof.

COROLLARY 4: *With* $\beta := 1/\theta$, *we have*

$$\mathbf{E}(L_k) = \frac{(\beta+1)\Gamma(n)}{\Gamma(\beta+n+1)} \frac{\Gamma(\beta+n-k+1)}{\Gamma(n-k+1)}, \qquad k = 1,\ldots,n, \tag{16}$$

with $\sum_{k=1}^n \mathbf{E}(L_k) = 1$.

PROOF: Putting $q = 1$ in the expression of $\mathbf{E}[L_k^q]$ and if $\beta = 1/\theta$, we get

$$\begin{aligned}\mathbf{E}(L_k) &= \prod_{i=1}^{k-1} \frac{(n-i)\theta}{(n-i+1)\theta+1} \cdot \frac{1+\theta}{(n-k+1)\theta+1} \\ &= (\beta+1)\frac{(n-1)!}{(n-k)!} \frac{1}{\prod_{i=0}^{k-1}(\beta+n-i)} \\ &= (\beta+1)\frac{\Gamma(n)}{\Gamma(n-k+1)} \frac{\Gamma(\beta+n-k+1)}{\Gamma(\beta+n+1)}.\end{aligned}$$

From normalization, it holds by construction that $\sum_{k=1}^n \mathbf{E}(L_k) = 1$. ■

https://doi.org/10.1017/S0269964805050060 Published online by Cambridge University Press

Let us now compute the joint distribution of the size-biased permutation $\mathbf{L}_n$ of $\mathbf{S}_n$. We will say in the sequel that, if $\mathbf{L}_n = \mathrm{SBP}(\mathbf{S}_n)$, then $\mathbf{L}_n \overset{d}{\sim} SBD_n(\theta)$ assuming that $\mathbf{S}_n \overset{d}{\sim} D_n(\theta)$.

3.3. Joint Law of the SBP

Let us first discuss the visiting order of the fragments in the SBP process. For any permutation $(m_1, \ldots, m_n)$ of $(1, \ldots, n)$, with $M_1, \ldots, M_k$, $k = 1, \ldots, n$, the first k distinct fragments' numbers that have been visited in the SBP sampling process, we have

$$\mathbf{P}(M_1 = m_1, \ldots, M_k = m_k | \mathbf{S}_n) = \prod_{i=1}^{k-1} \frac{S_{m_i}}{1 - \sum_{l=1}^{i} S_{m_l}} S_{m_k} \tag{17}$$

so that

$$\mathbf{P}(M_k = m_k | \mathbf{S}_n, M_1 = m_1, \ldots, M_{k-1} = m_{k-1}) = \frac{S_{m_k}}{1 - \sum_{l=1}^{k-1} S_{m_l}}. \tag{18}$$

As a result,

$$\mathbf{P}(M_k = m | \mathbf{S}_n) = S_m \sum_{(m_1 \neq \ldots \neq m_{k-1}) \neq m} \prod_{i=1}^{k-1} \frac{S_{m_i}}{1 - \sum_{l=1}^{i} S_{m_l}} \tag{19}$$

is the probability that the kth visited fragment is fragment number m from $D_n(\theta)$. If K_m is the random position of fragment number m, we then clearly have

$$\mathbf{P}(K_m = k | \mathbf{S}_n) = \mathbf{P}(M_k = m | \mathbf{S}_n), \tag{20}$$

translating the fact that K_m and M_k are inverses of one another, hence with $K_{M_k} = k$ and $M_{K_m} = m$.

Let us now compute the joint distribution of the size-biased permutation $\mathbf{L}_n$ of $\mathbf{S}_n$ with $\mathbf{L}_n \overset{d}{\sim} SBD_n(\theta)$ and $\mathbf{S}_n \overset{d}{\sim} D_n(\theta)$. First, we have

$$(L_1, \ldots, L_n) = (S_{M_1}, \ldots, S_{M_n}), \tag{21}$$

and, consequently,

$$\mathbf{P}(L_1 = S_{m_1}, \ldots, L_n = S_{m_n} | \mathbf{S}_n) = \prod_{k=1}^{n-1} \frac{S_{m_k}}{1 - \sum_{l=1}^{k} S_{m_l}} S_{m_n}, \tag{22}$$

the average of which over $\mathbf{S}_n$ gives the joint law of $\mathbf{L}_n := (L_1, \ldots, L_n)$.

We will now consider the joint moment function of the random size-biased permutation $\mathbf{L}_n = (L_1, \ldots, L_n)$. Indeed, we observe from Eqs. (12) and (13) and the independence of the V_k's that

$$\mathbf{E}\left[\prod_{k=1}^{n} L_k^{q_k}\right] = \mathbf{E}\left[\prod_{k=1}^{n} \prod_{i=1}^{k-1} \overline{V}_i^{q_k} V_k^{q_k}\right] = \prod_{k=1}^{n-1} \mathbf{E}\left[V_k^{q_k} \overline{V}_k^{q_{k+1}+\cdots+q_n}\right], \tag{23}$$

with $V_k \overset{d}{\sim} \text{beta}(1+\theta, (n-k)\theta)$ and $\overline{V}_k \overset{d}{\sim} \text{beta}((n-k)\theta, 1+\theta)$, $k = 1, \ldots, n-1$.

Putting all of this together, we obtain the following result.

THEOREM 5: *The joint moment function of the SBP* $\mathbf{L}_n = (L_1, \ldots, L_n) \overset{d}{\sim} SBD_n(\theta)$ *reads*

$$\mathbf{E}\left[\prod_{k=1}^{n} L_k^{q_k}\right] = \prod_{k=1}^{n-1} \left\{ \frac{\Gamma(1+(n-k+1)\theta)}{\Gamma(1+\theta)\Gamma((n-k)\theta)} \frac{\Gamma(1+\theta+q_k)\Gamma((n-k)\theta + q_{k+1} + \cdots + q_n)}{\Gamma(1+(n-k+1)\theta + q_k + \cdots + q_n)} \right\}. \tag{24}$$

PROOF: Let $V \overset{d}{\sim} \text{beta}(a,b)$. Then, with $\overline{V} := 1 - V$, it holds that

$$\begin{aligned}\mathbf{E}[V^{q_1}\overline{V}^{q_2}] &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 v^{a+q_1-1}(1-v)^{b+q_2-1}\,dv \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+q_1)\Gamma(b+q_2)}{\Gamma(a+b+q_1+q_2)}.\end{aligned}$$

Adapting this computation, recalling that $V_k \overset{d}{\sim} \text{beta}(1+\theta, (n-k)\theta)$, the quantity $\mathbf{E}[V_k^{q_k} \overline{V}_k^{q_{k+1}+\cdots+q_{n-1}}]$ has the expression displayed inside the product from Eq. (23). ■

Remark: Letting $q_k = q/n$, $k = 1, \ldots, n$, in Eq. (24), the moment function of the geometric average of $\mathbf{L}_n$, which is $\prod_{k=1}^{n} L_k^{1/n}$, follows.

4. COMPARATIVE SEARCH COST IN DIRICHLET PARTITION AND IN THE SIZE-BIASED PERMUTATION OF IT

We now show how these results can be used when considering an arcane problem from applied probability.

A collection of n books with random popularities S_m, $m = 1, \ldots, n$, is arranged on a shelf. (If instead of a collection of books, a population of n species were considered, popularities verbatim interpret as species abundance; see Kingman [15] and Ewens [7] for such interpretations).

Books' popularities are assumed to satisfy $\mathbf{S}_n \overset{d}{\sim} D_n(\theta)$. When a book is demanded, it is removed and replaced (before a next demand) to the top of the shelf, other books being shifted accordingly; successive demands are independent.

Iterating this heaps process (as a recurrent positive Markov chain over the set of permutations), there is intuitively a tendency when the system has reached equilibrium, to find more popular books to the top of the heap. At equilibrium indeed (see Donnelly [5] and references therein to the works of Dies, Hendricks, and Letac), books' popularities are given by $\mathbf{L}_n = \mathrm{SBP}(\mathbf{S}_n) \overset{d}{\sim} SBD_n(\theta)$ and result (iii) in Theorem 3 stating that $L_1 \geqslant \cdots \geqslant L_n$ confirms and gives some flesh to this intuition. Note from this that $\mathbf{L}_n = \mathrm{SBP}(\mathbf{L}_n)$ ($\mathbf{L}_n$ is invariant under size-biased permutation) and that $\mathbf{L}_n = \mathrm{SBP}(\mathbf{S}_{(n)})$ since $\mathbf{S}_{(n)}$ is simply obtained from $\mathbf{S}_n$ while rearranging its components in descending order.

Next, define the search cost of an item in a library to be the number of items above it in the heap; a weighted sum over the items yields the search cost of a typical item. The search cost when the library has reached the equilibrium state $\mathbf{L}_n$ is, of course, expected to be smaller than the search cost in $\mathbf{S}_n$ itself. We would like to revisit these ancient questions in the light of our preceding results on $\mathbf{L}_n = \mathrm{SBP}(\mathbf{S}_n)$ when $\mathbf{S}_n \overset{d}{\sim} D_n(\theta)$.

4.1. Search Cost in $\mathbf{S}_n \overset{d}{\sim} D_n(\theta)$

We start with computing the search cost $C_{n,\mathbf{S}}$ assuming popularities to be Dirichlet distributed. Here, $C_{n,\mathbf{S}}$ is the discrete random variable taking the value $m-1$ with probability $\mathbf{E}(S_m) = 1/n$, $m = 1,\ldots,n$. The moment generating function of $C_{n,\mathbf{S}}$ is expressed as

$$\mathbf{E}[e^{-\lambda C_{n,\mathbf{S}}}] = \frac{1}{n}\sum_{m=1}^{n} e^{-\lambda(m-1)} = \frac{1}{n}\frac{1-e^{-\lambda n}}{1-e^{-\lambda}}. \tag{25}$$

As a result, $\mathbf{E}(C_{n,\mathbf{S}}) = (n-1)/2$, $\mathbf{E}(C_{n,\mathbf{S}}^2) = (n-1)(2n-1)/6$, and $\boldsymbol{\sigma}^2(C_{n,\mathbf{S}}) = (n-1)(n+1)/12$, and we have the following proposition.

Proposition 6: *With U a uniformly distributed random variable on $(0,1)$, it holds that*

$$\frac{C_{n,\mathbf{S}}}{n} \xrightarrow{d} U \quad \textit{as } n \uparrow \infty. \tag{26}$$

Proof: From the expression of the moment generating function of $C_{n,\mathbf{S}}$, we have

$$\mathbf{E}[e^{-\lambda(C_{n,\mathbf{S}}/n)}] \xrightarrow{d} \frac{1-e^{-\lambda}}{\lambda},$$

which is the Laplace-Stieltjes transform of a uniformly distributed random variable U on $(0,1)$, with mean value $\frac{1}{2}$. Although in a different (deterministic) partition context, a similar result can be found in Fill [9, Thm. 4.2, p. 198]. ■

4.2. Search Cost in $\mathbf{L}_n = \mathrm{SBP}(\mathbf{S}_n) \overset{d}{\sim} SBD_n(\theta)$

From its definition, the search cost in $\mathbf{L}_n$ is the mixture $C_{n,\mathbf{L}} = K_M - 1$ (the number of fragments above fragment M in the list). Consequently, given $\mathbf{S}_n$, $C_{n,\mathbf{L}}$ will take the value $K_m - 1$ with probability S_m. Its conditional distribution is

$$\mathbf{P}(C_{n,\mathbf{L}} = k-1 | \mathbf{S}_n) = \sum_{m=1}^{n} \mathbf{P}(K_m = k | \mathbf{S}_n) S_m, \qquad k = 1, \dots, n,$$

where $\mathbf{P}(K_m = k | \mathbf{S}_n)$ is given by Eqs. (19) and (20). Let us first recall some well-known results on conditional search cost in $\mathbf{L}_n$, as a functional of $\mathbf{S}_n$.

When $\mathbf{P}(K_m = k | \mathbf{S}_n)$ takes the more usual form

$$\mathbf{P}(K_m = k | \mathbf{S}_n) = S_m \sum_{l=0}^{k-1} (-1)^{k-1-l} \binom{n-1-l}{k-1-l} \sum_{|J|=l;\, m \notin J} (1 - S_J)^{-1} \tag{27}$$

with $S_J = \sum_{j \in J} S_j$, the average position of original fragment m in the limiting partition $SBD_n(\theta)$ is known to be

$$\mathbf{E}(K_m | \mathbf{S}_n) = \sum_{k=1}^{n} k \mathbf{P}(K_m = k | \mathbf{S}_n) = 1 + \sum_{l \neq m} \frac{S_l}{S_l + S_m}, \tag{28}$$

so that the expected search cost in a $SBD_n(\theta)$ partition is

$$\mathbf{E}(C_{n,\mathbf{L}} | \mathbf{S}_n) := \sum_{m=1}^{n} S_m \{\mathbf{E}(K_m | \mathbf{S}_n) - 1\} = 2 \sum_{l<m} \frac{S_l S_m}{S_l + S_m}. \tag{29}$$

The results [Eqs. (27)–(29)] were obtained by Burville and Kingman [2]. They are valid for any random (or not) partition $\mathbf{S}_n$. Using Poisson embedding techniques, Fill and Holst [10], following combinatorial results of Flajolet, Gardy, and Thimonier [11], also found the full conditional generating function of $C_{n,\mathbf{L}}$ under the form

$$\mathbf{E}(u^{C_{n,\mathbf{L}}} | \mathbf{S}_n) = \int_0^\infty e^{-t} \left[\sum_{m=1}^{n} \frac{S_m^2}{1 + (e^{S_m t} - 1) u} \right] \left[\prod_{m=1}^{n} (1 + (e^{S_m t} - 1) u) \right] dt. \tag{30}$$

See also Hildebrand [13] for related problems.

Averaging over $\mathbf{S}_n$ gives the search-cost distribution in the random partition case. This is not so simple a matter, as it involves complicated Dirichlet integrals as it stands, in general. When $\mathbf{S}_n \overset{d}{\sim} D_n(\theta)$, averaging over $\mathbf{S}_n$, Kingman [14] obtained $\mathbf{E}(C_{n,\mathbf{L}}) = (n-1)\theta/(2\theta + 1)$. Recently, Barrera and Paroissin [1, Thm. 1] gave the full generating function $\mathbf{E}(u^{C_{n,\mathbf{L}}})$ under the form of a not so explicit double integral, even in the particular Dirichlet example.

Using the results of the preceding section, we will now show that the full law of $C_{n,\mathbf{L}}$ can be computed more explicitly when $\mathbf{S}_n \overset{d}{\sim} D_n(\theta)$ and $\mathbf{L}_n = \mathrm{SBP}(\mathbf{S}_n) \overset{d}{\sim} SBD_n(\theta)$. Several conclusions can next be drawn in this particular case.

Stated differently, indeed, $C_{n,\mathbf{L}}$ takes the value $k-1$ with probability $L_k = S_{M_k}$, $k = 1, \dots, n$, recalling that M_k and K_m are inverses of one another. Averaging over

https://doi.org/10.1017/S0269964805050060 Published online by Cambridge University Press

$\mathbf{L}_n$, $C_{n,\mathbf{L}}$ therefore takes the value $k-1$ with probability $\mathbf{E}(L_k)$. As a result, with $\beta := 1/\theta$ the inverse of temperature (or disorder), we obtain the following lemma.

LEMMA 7:

(i) *The law of* $C_{n,\mathbf{L}}$ *is given by*

$$\mathbf{P}(C_{n,\mathbf{L}} = k) = \frac{(\beta+1)\Gamma(n)}{\Gamma(\beta+n+1)} \frac{\Gamma(\beta+n-k)}{\Gamma(n-k)}, \qquad k = 0,\ldots,n-1; \quad \textbf{(31)}$$

it is unimodal, with mode at $k = 0$.

(ii) *The first and second moments are*

$$\mathbf{E}(C_{n,\mathbf{L}}) = \frac{n-1}{2+\beta} \quad \textit{and} \quad \mathbf{E}[C_{n,\mathbf{L}}^2] = \frac{(n-1)(2n+\beta-1)}{(\beta+2)(\beta+3)}. \qquad \textbf{(32)}$$

PROOF:

(i) The first part is a consequence of the explicit expression of $\mathbf{E}(L_k)$, $k = 1,\ldots,n$, appearing in Corollary 4. Note that $\mathbf{P}(C_{n,\mathbf{L}} = 0) = (\beta+1)/(\beta+n) = (1+\theta)/(1+n\theta)$, which is $\mathbf{E}(L_1)$. Next, for each $k = 0,\ldots,n-1$, we have

$$\frac{\mathbf{P}(C_{n,\mathbf{L}} = k+1)}{\mathbf{P}(C_{n,\mathbf{L}} = k)} = \frac{n-k-1}{\beta+n-k-1} < 1$$

and the mode of this distribution is, thus, at $k = 0$.

(ii) Equivalently, the moment function of $C_{n,\mathbf{L}}$ is

$$\mathbf{E}[C_{n,\mathbf{L}}^q] = \sum_{k=1}^{n-1} k^q \mathbf{E}(L_{k+1}) = \frac{(\beta+1)\Gamma(n)}{\Gamma(\beta+n+1)} \sum_{k=1}^{n-1} k^q \frac{\Gamma(\beta+n-k)}{\Gamma(n-k)}$$
$$= \frac{(\beta+1)\Gamma(n)}{\Gamma(\beta+n+1)} \sum_{p=1}^{n-1} (n-p)^q \frac{\Gamma(\beta+p)}{\Gamma(p)}.$$

Putting $A := \Gamma(\beta+n+1)/[(\beta+1)\Gamma(n)]$, the normalization $\sum_{k=1}^{n} \mathbf{E}(L_k) = 1$ (with $q = 0$) reads

$$A = \sum_{p=1}^{n} \frac{\Gamma(\beta+p)}{\Gamma(p)}.$$

Let B be obtained from A while substituting $\beta+1$ to β and $n-1$ to n. We get

$$B = \frac{\Gamma(\beta+n+1)}{(\beta+2)\Gamma(n-1)} = \frac{(n-1)(\beta+1)}{\beta+2} A$$
$$= \sum_{p=1}^{n-1} \frac{\Gamma(\beta+p+1)}{\Gamma(p)} = \sum_{p=1}^{n-1} p \frac{\Gamma(\beta+p+1)}{\Gamma(p+1)}$$
$$= \sum_{p=2}^{n} (p-1) \frac{\Gamma(\beta+p)}{\Gamma(p)} = \sum_{p=1}^{n} p \frac{\Gamma(\beta+p)}{\Gamma(p)} - A.$$

https://doi.org/10.1017/S0269964805050060 Published online by Cambridge University Press

Putting $q = 1$ in $\mathbf{E}[C_{n,\mathbf{L}}^q]$, we obtain

$$\begin{aligned}\mathbf{E}(C_{n,\mathbf{L}}) &= \frac{1}{A}\left(n\sum_{p=1}^{n-1}\frac{\Gamma(\beta+p)}{\Gamma(p)} - \sum_{p=1}^{n-1} p\,\frac{\Gamma(\beta+p)}{\Gamma(p)}\right)\\ &= \frac{1}{A}\left[n\left(A - \frac{\Gamma(\beta+n)}{\Gamma(n)}\right) - \left(A + B - n\,\frac{\Gamma(\beta+n)}{\Gamma(n)}\right)\right]\\ &= \frac{1}{A}\left[nA - A\left(1 + \frac{(n-1)(\beta+1)}{\beta+2}\right)\right]\\ &= \frac{n-1}{2+\beta}.\end{aligned}$$

Using similar arguments, the second-order moment ($q = 2$) can be computed. In more detail, if C is obtained from A while substituting $\beta + 2$ to β and $n - 2$ to n, we get

$$C = \frac{(n-1)(n-2)(\beta+1)}{\beta+3}A = \sum_{p=1}^{n} p^2\,\frac{\Gamma(\beta+p)}{\Gamma(p)} - 3B - A.$$

Putting $q = 2$ in $\mathbf{E}[C_{n,\mathbf{L}}^q]$, we obtain

$$\begin{aligned}\mathbf{E}(C_{n,\mathbf{L}}^2) &= \frac{1}{A}\,(n^2A - 2n(A+B) + C + 3B + A)\\ &= \frac{(n-1)(2n+\beta-1)}{(\beta+2)(\beta+3)}.\end{aligned}$$

The result on the mean value was obtained by Kingman [14], with different techniques. Note that $\mathbf{E}(C_{n,\mathbf{S}}) \geq \mathbf{E}(C_{n,\mathbf{L}})$, as expected. The result on the variance (with a different proof) is also in Barrera and Paroissin's example 2 [1]. In fact, the full preasymptotic law of $C_{n,\mathbf{L}}$ is available from part (i), which seems to be new. ■

From our approach, we also obtain the asymptotic result.

THEOREM 8: *As $n \uparrow \infty$,*

$$\frac{C_{n,\mathbf{L}}}{n} \xrightarrow{d} B_{1,1+1/\theta}, \tag{33}$$

where $B_{1,1+1/\theta} \overset{d}{\sim} beta(1,1+1/\theta)$.

PROOF: From the expression of the moment function of $C_{n,\mathbf{L}}$, we have

$$\mathbf{E}\left[\left(\frac{C_{n,\mathbf{L}}}{n}\right)^q\right] = \frac{(\beta+1)\Gamma(n)}{\Gamma(\beta+n+1)}\sum_{p=1}^{n-1}\left(1-\frac{p}{n}\right)^q\frac{\Gamma(\beta+p)}{\Gamma(p)}.$$

https://doi.org/10.1017/S0269964805050060 Published online by Cambridge University Press

For large n, this can be approximated by

$$\mathbf{E}\left[\left(\frac{C_{n,\mathbf{L}}}{n}\right)^q\right] \sim \frac{(\beta+1)\Gamma(n)}{\Gamma(\beta+n)} \int_0^1 (1-x)^q \frac{\Gamma(\beta+nx)}{\Gamma(nx)}\, dx.$$

From Stirling's formula, with $a > 0$, $\Gamma(a+z)/\Gamma(z) \sim_{z\uparrow\infty} z^a$. This shows that for large n,

$$\begin{aligned}\mathbf{E}\left[\left(\frac{C_{n,\mathbf{L}}}{n}\right)^q\right] &\sim \frac{(\beta+1)}{n^\beta} \int_0^1 (1-x)^q (nx)^\beta\, dx \\ &= (\beta+1)\int_0^1 (1-x)^q x^\beta\, dx \\ &= (\beta+1)\int_0^1 x^q (1-x)^\beta\, dx,\end{aligned}$$

which is the moment function of a beta$(1,1+\beta)$ random variable with mean value $1/(2+\beta) = \theta/(2\theta+1) < \frac{1}{2}$. ■

Remark: The limiting search cost per item in $\mathbf{L}_n = \mathrm{SBP}(\mathbf{S}_n) \overset{d}{\sim} SBD_n(\theta)$ is distributed like $B_{1,1+1/\theta}$, whereas its law is the one of a uniform random variable U in $\mathbf{S}_n \overset{d}{\sim} D_n(\theta)$. Clearly, we have $U \geqslant B_{1,1+1/\theta}$, expressing the fact that search-cost per item from the random partition $D_n(\theta)$ is asymptotically larger than from the $SBD_n(\theta)$ one, which is more organized. This result differs from the well-known one that $\mathbf{E}(C_{n,\mathbf{S}}) \geq \mathbf{E}(C_{n,\mathbf{L}})$ for each n.

4.3. Search Cost in the Kingman Limit

Consider the situation where $n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$. Such an asymptotics was first considered by Kingman. When $k = o(n)$, recalling $V_k \overset{d}{\sim} \mathrm{beta}(1+\theta, (n-k)\theta)$, we have $V_k \xrightarrow{d} U_k \overset{d}{\sim} \mathrm{beta}(1,\gamma)$ and the $SBD_n(\theta)$ distribution converges weakly to a GEM(γ) distribution (GEM = Griffiths-Engen-McCloskey). Namely, $(L_1,\ldots,L_n) \xrightarrow{d} (L_1^*,\ldots,L_k^*,\ldots) =: \mathbf{L}^*$, where

$$L_k^* = \prod_{i=1}^{k-1} \bar{U}_i U_k, \qquad k \geq 1. \tag{34}$$

Here, $(U_k, k \geq 1)$ are independent and identically distributed with law $U_1 \overset{d}{\sim} \mathrm{beta}(1,\gamma)$ and $\bar{U}_1 := 1 - U_1 \overset{d}{\sim} \mathrm{beta}(\gamma,1)$. Note that $L_1^* \geqslant \cdots \geqslant L_k^* \geqslant \cdots$ and that $\mathbf{L}^*$ is invariant under size-biased permutation. In the Kingman limit, $(S_{(m)}, m = 1,\ldots,n)$ converges in law to a Poisson–Dirichlet distribution $(L_{(k)}^*, k \geq 1) \overset{d}{\sim} \mathrm{PD}(\gamma)$ with $L_{(1)}^* > \cdots > L_{(k)}^* > \cdots$. The size-biased permutation of $(L_{(k)}^*, k \geq 1)$ is $(L_k^*, k \geq 1) \overset{d}{\sim} \mathrm{GEM}(\gamma)$ (see Kingman [16, Chap. 9], Tavaré and Ewens [18], and references to Pitman's work therein for a review). As a result, we have the following:

https://doi.org/10.1017/S0269964805050060 Published online by Cambridge University Press

PROPOSITION 9: *As* $n \uparrow \infty$, $\theta \downarrow 0$ *while* $n\theta = \gamma > 0$,

$$C_{n,\mathbf{L}} \xrightarrow{d} C_{\mathbf{L}^*} \overset{d}{\sim} \mathrm{geom}(\gamma). \tag{35}$$

PROOF: In particular, we have $\mathbf{E}(L_k^*) = [\gamma/(1+\gamma)]^{k-1}[1/(1+\gamma)]$. The moment generating function of the search cost in the Kingman limit is thus

$$\mathbf{E}[e^{-\lambda C_{\mathbf{L}^*}}] = \sum_{k\geq 1} e^{-\lambda(k-1)}\mathbf{E}(L_k^*) = \frac{1}{1+\gamma(1-e^{-\lambda})},$$

which is the one of a geometric distribution with mean γ.

Note that $\mathbf{P}(C_{\mathbf{L}^*} = 0) = 1/(1+\gamma)$, which is the average length of L_1^*. ■

References

1. Barrera, J. & Paroissin, C. (2004). On the distribution of the stationary search cost for the move-to-front rule with random weights. *Journal of Applied Probability* 41(1): 250–262.
2. Burville, P.J. & Kingman, J.F.C. (1973). On a model of storage and search. *Journal of Applied Probability* 10: 697–701.
3. Collet, P., Huillet, T., & Martinez, S. (2003). Size-biased picking for finite random partitions of the interval. Preprint; *Journal of Applied Probability* (submitted).
4. Donnelly, P. (1986). Partition structures, Pòlya urns, the Ewens sampling formula and the age of alleles. *Theoretical Population Biology* 30: 271–288.
5. Donnelly, P. (1991). The heaps process, libraries and size-biased permutation. *Journal of Applied Probability* 28: 321–335.
6. Engen, S. (1978). *Stochastic abundance models*. Monographs on Applied Probability and Statistics. London: Chapman & Hall.
7. Ewens, W.J. (1990). Population genetics theory—the past and the future. In S. Lessard (ed.), *Mathematical and statistical developments of evolutionary theory*. Dordrecht: Kluwer.
8. Feller, W. (1971). *An introduction to probability theory and its applications*, Vol. 2, 2nd ed. New York: Wiley.
9. Fill, J.A. (1996). Limits and rates of convergence for the distribution of search cost under the move-to-front rule. *Theoretical Computer Science* 164: 185–206.
10. Fill, J. A. & Holst, L. (1996). On the distribution of search cost for the move-to-front rule. *Random Structures and Algorithms* 8(3): 179–186.
11. Flajolet, P., Gardy, D., & Thimonier, L. (1992). Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Applied Mathematics* 39: 207–229.
12. Hawkes, J. (1981). On the asymptotic behaviour of sample spacings. *Mathematical Proceedings of the Cambridge Philosophical Society* 90(2): 293–303.
13. Hildebrand, M. (1999). On a conjecture of Fill and Holst involving the move-to-front rule and cache faults. *Probability in the Engineering and Informational Sciences* 13(3): 377–385.
14. Kingman, J.F.C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society. Series B* 37: 1–22.
15. Kingman, J.F.C. (1978). Random partitions in population genetics. *Proceedings of the Royal Society London Series A* 361(1704): 1–20.
16. Kingman, J.F.C. (1993). *Poisson processes*. Oxford: Clarendon Press.
17. Patil, G.P. & Taillie, C. (1977). Diversity as a concept and its implications for random environments. *Bulletin de l'Institut International de Statistique* 4: 497–515.
18. Tavaré, S. & Ewens, W.J. (1997). Multivariate Ewens distribution. In N.L. Johnson, S. Kotz, & N. Balakrishnan (eds.), *Discrete multivariate distributions*. New York: Wiley, pp. 232–246.

https://doi.org/10.1017/S0269964805050060 Published online by Cambridge University Press